Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 152 – 158

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Using Weighted Model Averaging in Distributed Multilingual DNNs to Improve Low Resource ASR

Reza Sahraeian, Dirk Van Compernolle*

KU Leuven - ESAT, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

Abstract

Multilingual Deep Neural Networks (DNNs) have been successfully used to leverage out-of-language data to boost the performance of a low resource ASR. However, the mismatch between auxiliary source languages and the target language can leave a negative effect on acoustic modeling for the target language. Thus, a key challenge in multilingual DNNs is to exploit acoustic data from multiple donor languages to improve on ASR performance while mitigating the problem of language mismatch. In this paper, we propose to employ weighted model averaging in the framework of distributed multilingual DNN which allows the target language or similar languages to take higher weights during the multilingual DNN training, and consequently shift the parameters towards the acoustic space of target data. Furthermore, we utilize the same strategy in the adaptation phase where a conventional multilingual DNN is the starting point and retraining is applied using all languages with different weights. The experiments with four languages from the GlobalPhone dataset show that the recognition performances in both scenarios are improved. The latter, moreover, provides a low-cost and efficient methodology for multilingual DNNs.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

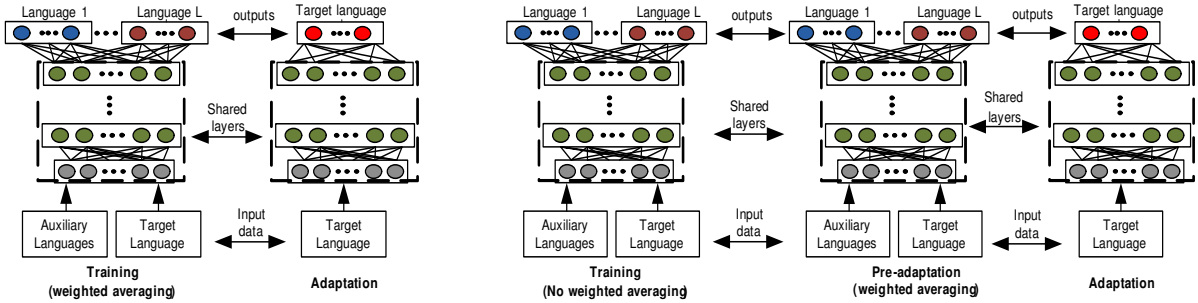
Keywords: Low resource ASR; multilingual DNN; distributed DNN training

1. Introduction

In the field of low resource speech recognition, the prevailing paradigm is to take advantage of external *source languages* mostly in the framework of multilingual deep neural networks [1,2,3]. Multilingual DNNs are used either based on the hybrid [3] or tandem systems [4], and they have provided significant improvements in various speech recognition tasks [5,6]. In multilingual DNNs, transferring knowledge across languages is accomplished by sharing the parameters of the hidden layers. In other words, the hidden layers are considered as universal feature extractors while the softmax layers are language dependent. Additional improvement can be obtained by further adjusting the whole DNN which is often termed as DNN *adaptation* [7,8].

* Corresponding author. Tel.: +32-16-321723 ; fax: +32-16-321055.
E-mail address: {reza.sahraeian,dirk.vancompernolle}@kuleuven.be

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.043



(a) Using weighted model averaging in training.

(b) Utlizing weighted model averaging as a pre-adaptation phase.

Fig. 1: Two types of multilingual DNN training in which weighted model averaging can be employed in the training (a) or pre-adaptation (b).

The performance of multilingual DNNs, however, can be affected by two opposing factors. From one point of view, using larger amount of training data and various source languages increases the chance of having more generalized multilingual DNN with better context coverage. On the other hand, the difference between the target language and the source language(s) may obtrude with impurification of training data and hurt target language's acoustic model[9,10]. Especially, when reasonable amount of training data for target language is available, the negative effect of language mismatch may even make multilingual DNN performs worse than the monolingual system. To mitigate this problem, it is proposed to use a language identification approach to pick the language(s) most similar to the target language from a given set of source languages[11]; it is shown that the multilingual DNN trained on the similar language(s) outperforms the one trained on all available source languages. Furthermore, a set of experiments are provided to investigate whether it is better to utilize data from similar languages or more data from diverse languages in the multilingual DNN training[12]. It is shown that when multilingual DNN training employs "best fitting" languages, significant improvement is obtained; moreover, even adding mismatched languages gives gains over the monolingual baseline if the set of source languages is big enough. Such techniques suggest to utilize different multilingual training data for a new target language which is not of interest because of the following reasons: first, since the DNN configuration depends on the amount of training data, we need to tune the number of hidden layers and neurons again which is not practically efficient. Moreover, multilingual DNN can still benefit from not closely related languages[12] and thus removing them from the training set is the naive approach. This motivates us to investigate a method in which all languages are incorporated in the multilingual DNN while the more important languages play dominant roles.

In this work, we employ a weighted average modeling approach in the framework of distributed multilingual DNN training. Parallelized training of DNN acoustic models has been a popular topic in recent years in both monolingual[13,14,15] and multilingual[16,17] scenarios to accelerate large networks training. In our study, however, we utilize the parallel training and model combination by proposing a weighted model averaging to improve the performance of a multilingual DNN. In this framework, different languages can take different weights and thus leave different effects on the multilingual DNN. First, we investigate the proposed method during multilingual DNN training and we demonstrate its usefulness. Furthermore, we employ our method as a *pre-adaptation* technique in which a conventional multilingual DNN is retrained using the weighted model averaging approach and then the resultant model will be adapted using only target language data. Finally, we properly integrate our recently proposed rank-constrained multilingual DNN approach[18] with the proposed pre-adaptation technique to obtain further improvements.

The remainder of this paper is organized as follows: In section 2, we describe the distributed multilingual DNN training method and weighted model averaging. The experimental setup and results are presented in sections 3 and 4. Finally, we have concluding remarks.

2. Multilingual DNN

A conventional multilingual DNN with shared hidden layers has been formulated in the multi-task learning framework. The hidden layers as feature extractors and the classifiers are jointly optimized on the shared data for different languages. In other words, data from all training languages are fed through the common hidden layers but each language has its own output layer[3]. The output layers are usually context-dependent states determined by standard clustering algorithms from previously trained HMMs. The most commonly used optimization procedure for DNN training is stochastic gradient descent (SGD)[19] which is a sequential training method and makes scaling neural networks difficult. Therefore, different model and data parallelization training schemes have been investigated in the literature to speed up DNN training.

In this study, we use the methodology based on distributed model averaging[15]. This method uses a version of data parallelism which allows multiple SGD being processed on different machines and the model parameters are averaged across all machines after a fixed number of samples has been processed. In this work, we set this number of samples to 400000[15]. The averaged parameters are then redistributed for the next iteration and it will be repeated until all the data are processed for a specific number of epochs. Moreover, DNNs are trained based on greedy layerwise supervised training[20].

The same framework can be utilized for multilingual DNN training. To this end, we use the language-based distributed learning algorithm in which each GPU uses the full data from one language and trains the normal DNN model. For the sake of efficiency, we consider the same number of samples, which is 400000, to be processed for each language before averaging the parameters. This reduces the waiting time before averaging, but we need to consider different epochs for languages depending on the amount of available training data for each language. Moreover, we only average the parameters of input and hidden layers across languages and keep the output layers language dependent. Although the initialization of the multilingual DNN can be done in the greedy layerwise manner, in this work, we simply borrow the hidden layers from an already trained DNN and only randomly initialize the softmax layer.

2.1. Weighted model averaging

Instead of simple averaging, we propose to calculate the weighted average of the models. This allows us to control the effect of different languages on the multilingual DNN. Let's assume there are L languages being used for multilingual DNN training; then, we have L models being learned in parallel: $\{\Gamma_1, ..., \Gamma_L\}$. The parameters for model l can be represented by $\Lambda_l \equiv \{\Lambda^l_{shared} \cup \Lambda^l_{output}\}$. Λ^l_{output} refers to the language specific parameters in the softmax layer for language l and Λ^l_{shared} consists of model parameters that are supposed to be shared across all languages; thus, weighted averaging is applied on these parameters:

$$\Lambda_{shared} = \sum_{l=1}^{L} \lambda_l \Lambda^l_{shared} \tag{1}$$

where λ_l is the corresponding weight for language l and $\sum_{l=1}^{L} \lambda_l = 1$. This occurs periodically after 400000 samples from each language is processed and the resultant parameters are redistributed as the starting point for further training. The key to successful application of this approach is to properly choose λs for the training languages. For example, if data from the target language exists in the training set, it makes sense to give a higher weight to the parameters being trained over the target language data. This idea can be extended such that similar languages to the target language also take higher weights. Specifically in the case that only a small amount of target language data is available, the higher weights for the parameters being trained over matched languages may improve the acoustic model for the target language.

The multilingual DNN parameters being trained in the weighted averaging framework (Fig. 1 (a)), however, are shifted towards the languages with higher weights and thus hidden layers cannot be considered as language independent feature extractors. That is, for a new target language, a new multilingual DNN should be trained which is not of interest due to the slow DNN training procedure. Besides, the commonly used adaptation method through retraining the network with language specific data may not succeed specially if the network is too big and/or adaptation data is small. To alleviate these problems, we also propose to use the weighted averaging in the adaptation phase where an

intermediate retraining phase is applied after multilingual DNN training and before adaptation with target language data as shown in Fig. 1 (b). This phase, which we call pre-adaptation phase, consists of retraining the multilingual DNN, which is already trained in a common way, for a small number of iterations using all languages in the weighted model averaging framework. It is clear that for a new target language, only the pre-adaptation and adaptation procedure need to be done which are much more faster than training a new multilingual DNN using the weighted model averaging from scratch.

It is worth noting that in this paper, we provide an illustrative study on the effect of the proposed model combination rather than a closed form solution to find λs.

3. Experimental setup

3.1. ASR systems

Monolingual reference systems were built using target language data only. First, Gaussian mixture model systems were built using 39-dimensional MFCC feature vectors with 13 cepstral coefficients, and their first and second derivatives. Speaker based cepstral mean and variance normalization (CMVN) was applied and features were spliced in time taking a context size of 7 frames (i.e.,$\pm$3), followed by decorrelation and dimensionality reduction to 40 using LDA and further decorrelation using MLLT[21]. The number of gaussians and tied states for GMM based modeling was tuned over the development set. The derived states were used as targets in the DNN systems.

Then, monolingual DNNs were trained on mean and variance normalized 23-dimensional FBANK features being concatenated with 7 left and 7 right neighbor frames to yield an input feature vector size of 345; we observed that FBANK features outperform MFCCs as input features for DNN. The input features and the learning rates for the multilingual DNNs were the same as those used in the monolingual DNNs except that normalization was not applied. More details about the implementations are provided in the experiment section.

All the DNNs used in this study were trained using a ReLU nonlinearity based on greedy layerwise supervised training[15]. The initial and final learning rates were specified by hand and equal to 0.01 and 0.001 respectively.

The Kaldi ASR toolkit[22] is used for both GMM and DNN based acoustic modeling.

3.2. Database: The GlobalPhone

The GlobalPhone corpus is a multilingual text and speech corpus that covers speech data from 20 languages[23]. In our experiments, German (GE) was used as the target language, and Arabic (AR), Turkish (TU) and French (FR) as the auxiliary languages. The detailed statistics for these languages from the Globalphone corpus are presented in [23]. The recognition task is a standard word recognition task using a trigram language model obtained from Karlsruhe University [1].

The full German database consists of 14.85 hours by 65 speakers. To simulate a very low resource condition, we also constructed a subset containing 1 hour (8 speakers) of data, using randomly selected 7-8 minutes of speech for each of the selected speakers. The development and evaluation set include 1.95 and 1.45 hour data and each of them consists of 6 speakers. For the multilingual experiments, we used respectively following amounts of data of the donor languages: 22.74hr for FR, 16.54hr for AR and 13.23hr for TU.

4. Experiments

First, continuous speech recognition is performed in a monolingual fashion. In this set of experiments, both HMM/GMM and HMM/DNN systems were used as explained in section 3.1. The number of context-dependent triphone states were 700 and 3100 with an average of 4 and 13 Gaussian components per state for 1hr and 14.85hr German training data respectively. These parameters were tuned on the development set. The development set was also used to tune the number of hidden layers and neurons per layer in the HMM/DNN system. The optimal number of

[1] http://csl.ira.uka.de/GlobalPhone/

Table 1: WER(%) for German using monolingual and multilingual DNN systems.

Settings		Monolingual		Multilingual DNN
		GMM	DNN	
1hr	Dev.	22.84	21.41	18.91
	Eval.	35.38	34.90	33.77
14.85hr	Dev.	13.95	11.56	11.26
	Eval.	21.36	19.49	18.28

Table 2: Averaged log-likelihood of data from different languages given the UBM for the 14.85hr German data scenario.

UBM	Averaged log-likelihood			
	GE	FR	TU	AR
GE	-90.69	-107.08	-108.09	-112.93

hidden layers were 4, 5 and the number of hidden units in each layer were 50 and 300 for 1hr and 14.85hr of training data respectively. Word error rates (WER) for both development (Dev.) and evaluation (Eval.) sets are summarized in Table 1 for HMM/GMM systems as well as HMM/DNN ones.

For our first multilingual experiments, we used the conventional multilingual DNN with a dedicated softmax layer for each language while the hidden and input layers were shared. The hidden layers are initialized with FR which as we will show later is the closest language to German. Following the setup of the authors in [1], the number of target context-dependent states were set to 3100 for each auxiliary language. We used a DNN with 7 layers for the setting including 1hr of German data and 8 layers for the other setting; the number of nodes was 1500 per layer in all DNNs. The performance of the multilingual systems with adaptation is presented in Table 1. From Table 1 we can observe that multilingual DNN performs the best.

In the next set of our experiments, we investigated if some languages may play more important roles than others during the multilingual DNN training. Therefore, we first set out a simple data driven approach to find the closeness between the auxiliary languages and the target language. To that end, we trained a Universal Background Model (UBM) with 400 Gaussian components using 39-dimensional MFCC features using German and then given this UBM, the log-likelihoods of other languages data were calculated. Table 2 shows these values for the setting with 14.85hr of German data; it suggests that FR is the closest language to GE while AR is the furthest. We also observed the same order of closeness by using a UBM trained with 1hr German.

Then, we assessed the performance of our proposed weighted model averaging approach in the training phase. Throughout the experiments in this part, we considered the ratios of weights for various languages rather than assigning specific values to them. For example, the weight ratio set of $(1:1:1:1)$ refers to a scenario that all languages take the same weight which is 0.25. First, we only adjusted the weight of the target language and assigned the same weights to the auxiliary languages as shown in the first four rows of the Table 3. The results reveal that giving a higher weight to the target language during the multilingual DNN training may improve the recognition performance. This is not surprising as the closest language to German is itself! However, we observe that in the case of having only 1hr of training data from German, the performance is degraded when the weight assigned to German is five times larger than other weights. We attribute this behavior to the fact that the DNN trained with such small amount of data is not reliable enough to take a very large weight. Thus, choosing a proper weight not only depends on the closeness of the corresponding language but also on the amount of training data available for that language.

Furthermore, we investigated scenarios where source languages also take different weights. The last two rows in Table 3 show the cases where the weight assigned to FR is bigger than AR and TU as FR shows the highest similarity to German based on Table 2. In the setting with 1hr German training data, no gain is obtained over the baseline multilingual system performance presented in Table 1. In the other setting with 14.85hr German, improvements are obtained compared to the baseline multilingual system; however, we can observe that the recognition performance when FR has the same weight as TU and AR is the best. We also conducted some other experiments with different combination of weight ratios and the same observation was always made. It seems that in our setting giving different weights to the auxiliary languages during the training cannot benefit the target language's acoustic model.

Table 3: WER(%) for German using different weight ratios in the weighted model averaging approach for multilingual DNN.

Language weight ratio (FR:TU:AR:GE)	German data			
	1hr		14.85hr	
	Dev	Eval	Dev	Eval
(1:1:1:1)	18.91	33.77	11.26	18.28
(1:1:1:2)	18.62	33.66	11.10	18.04
(1:1:1:3)	18.72	33.53	11.07	17.87
(1:1:1:5)	18.87	33.82	11.00	17.74
(2:1:1:3)	18.91	34.12	11.09	18.07
(2:1:1:5)	18.95	34.23	11.08	18.07

Table 4: WER(%) for German using weighted model averaging in the adaptation phase.

Language weight ratio (FR:TU:AR:GE)	German data			
	1hr		14.85hr	
	Dev	Eval	Dev	Eval
(1:1:1:2)	18.52	33.13	10.96	17.96
(1:1:1:3)	18.38	32.87	11.01	18.18
(1:1:1:5)	18.66	32.64	11.01	17.84
(2:1:1:5)	17.89	33.04	11.04	17.99

Next, we used the weighted model averaging scheme in the adaptation phase. To this end, the conventional multilingual DNN was used as a starting point for further training with the proposed method. Then, the resultant model was fine tuned via the training German data. Table 4 shows the WERs for some weight ratio combinations. The results show that the improvements achieved is comparable with those presented in Table 3. Another interesting trend is that in this set of experiments a small number of epochs, (one or two), was required for the pre-adaptation phase which makes it much more efficient and faster than using weighted model averaging in the training. The proper number of epochs for pre-adaptation and adaptation phases were set using the development set. Moreover, unlike what we observed in Table 3, in the case FR takes a higher weight than AR and TU the performances are improved in both settings. This confirms our hypothesis that giving a higher weights to the similar languages can be beneficial for target language acoustic model.

Finally, we also incorporated our recently proposed method by applying singular value decomposition on all weight layers except the input layer. This led to a sparsified network with a much smaller number of parameters; for example, in our experiments, each hidden weight layer consists of 1500×1500 parameters which can be factorized into two smaller matrices such that $(1500 \times n_r + n_r \times 1500) < 1500 \times 1500$ where n_r is the corresponding rank. This boosts the performance specifically in the adaptation phase where the small number of parameters need to be fine tuned with low resource target language data. We need to fine tune the network after the low rank factorization (LRF); thus, we can merge this fine tuning with the weighted model averaging scheme. In this study, we examined the use of LRF where $n_r = 500$ and the weighted model averaging is applied on top of the factorized network and adaptation with only German data is utilized at the end. We considered the weight ratio of $(2 : 1 : 1 : 5)$ which performs well as shown in Table 4. The results are shown in Table 5 and significant improvements are achieved compared to the conventional multilingual DNN.

5. Conclusions

We have demonstrated in this work that exploiting out-of-language data in the framework of multilingual DNNs can be boosted for a specific low resource target language by weighting the involved languages. We used weighted averaging of parameters in a distributed learning manner. We investigated this methodology in both training and

Table 5: Results of applying low rank factorization together with weighted model averaging for fine tuning.

Settings	German data			
	1hr		14.85hr	
	Dev	Eval	Dev	Eval
WER (%)	**16.95**	**29.51**	**10.52**	**16.59**

adaptation. From the combined set of experiments we may draw following conclusions: first, we can boost the ASR performance for the target language by assigning a higher weight to the target language in the training phase; however, the optimal value for this weight also depends on the amount of available training data. Second, weighted model averaging can be utilized as a pre-adaptation phase by retraining an already trained multilingual DNN; then, the achieved model may be further fine tuned via the target data. Not only does it improve the results but also it provides a low-cost and fast implementation. Finally, we employed the low rank factorization of multilingual DNN together with the proposed method and showed these two techniques can be used jointly to gain further improvement.

References

1. Ghoshal A, Swietojanski P, Renals S. Multilingual training of deep neural networks, *in ICASSP* 2013, pp. 7319-7323.
2. Grezl F, Karafiat M, Janda M. Study of probabilistic and bottle-neck features in multilingual environment, *in ASRU* 2011, pp. 359-364.
3. Huang JT, Li J, Yu D, Deng L, Gong Y. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, *in ICASSP* 2013, pp. 7304-7308.
4. Vesely K, Karafit M, Grezl F, Janda M, Egorova E. The language-independent bottleneck features, *in Spoken Language Technology Workshop* 2012, pp. 336-341.
5. Gales MJF, Knill KM, Ragni A, Rath SP. Speech recognition and keyword spotting for low resource languages: babel project research at CUED, *in Spoken Language Technologies for Under-Resourced Languages* 2014, pp. 16-23.
6. Knill KM, Gales MJF, Rath SP, Woodland PC, Zhang C, Zhang S-X. Investigation of multilingual deep neural networks for spoken term detection, *in ASRU* 2013, pp. 138-143.
7. Grezl F, Karafiat M, Vesely K. Adaptation of multilingual stacked bottle-neck neural network structure for new language, *in ICASSP* 2014, pp. 7654-7658.
8. Thomas S, Seltzer ML, Church K, Hermansky H. Deep neural network features and semi-supervised training for low resource speech recognition, *in ICASSP* 2013, pp. 6704-6708.
9. Lin H, Deng L, Yu D, Gong Y, Acero A, Lee C. A study on multilingual acoustic modeling for large vocabulary ASR, *in ICASSP* 2009, pp. 4333-4336.
10. Vu NT, Schultz T. Multilingual multilayer perceptron for rapid language adaptation between and across language families, *in INTERSPEECH* 2013, pp. 515-519.
11. Zhang Y, Chuangsuwanich E, Glass J. Language ID-based training of multilingual stacked bottleneck features, *in INTERSPEECH* 2014, pp. 1-5.
12. Muller M, Stuker S, Sheikh Z, Metze F, Waibel A. Multilingual deep bottle neck features:a study on language selection and training techniques, *in International Workshop on Spoken Language Translation* 2014, pp. 257-264.
13. Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Senior A, Tucker P, Yang K, Le QV, et al. Large scale distributed deep networks, *in Advances in Neural Information Processing Systems* 2012, pp. 1223-1231.
14. Seide F, Fu H, Droppo J, Li G, Yu D. On parallelizability of stochastic gradient descent for speech DNNs, *in ICASSP* 2014, pp. 235-239.
15. Povey D, Zhang X, Khudanpur S. Parallel training of deep neural networks with natural gradient and parameter averaging, *arXiv:1410.7455* 2014.
16. Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato MA, Devin M, Dean J. Multilingual acoustic models using distributed deep neural networks, *in ICASSP* 2013, pp. 8619-8623.
17. Miao Y, Zhang H, Metze F. Distributed Learning of Multilingual DNN Feature Extractors using GPUs, *in INTERSPEECH* 2014, pp. 830-834.
18. Sahraeian R, Van Compernolle D. A study of rank-constrained multilingual DNNs for low-resource ASR, *in ICASSP* 2016, pp. 5420-5424.
19. Bottou L, Stochastic gradient learning in neural networks, *Proceedings of Neuro-Nmes*, **91** 1991.
20. Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks, *in INTERSPEECH* 2011, pp. 437-440.
21. Gales MJF, Semi-tied covariance matrices for hidden markov models, *IEEE Transactions on Speech and Audio Processing* **7** 1999, pp. 272-281.
22. Povey D, et al. The KALDI speech recognition toolkit, *in ASRU* 2011, pp. 1-4.
23. Schultz T, Vu NT, Schlippe T. Globalphone: A multilingual text & speech database in 20 languages, *in ICASSP* 2013, pp. 8126-8130.